International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 7 (2025) 13–25

Paper Type: Original Article

Multimodal Fake News Detection: A Survey of Text and Visual Content Integration Methods

Mahinda Zidan [1,*], **Ahmed Sleem** [2], **Ayman Nabil** [3] **and Mahmoud Othman** [1]

[1] Department of Computer Science, Faculty of Computers and Information Technology, Future University in Egypt, Cairo, 11835, Egypt.
Emails: Mahinda.zidan@fue.edu.eg; msamy@fue.edu.eg.

[2] Department of Computer Science, Faculty of Computers and Information, Tanta University, Tanta, 31527, Egypt; Ahmed.selim@ics.tanta.edu.

[3] Faculty of Computers Science, Misr International University, Cairo, 11800, Egypt; ayman.nabil@miuegypt.edu.eg.

Received: 30 Dec 2024 **Revised**: 28 Feb 2025 **Accepted**: 28 Apr 2025 **Published**: 30 Apr 2025

Abstract

The widespread use of social media has significantly facilitated the transmission of misleading information, making fake news a complicated and important global issue. While misinformation has emerged for decades, it has evolved from simple text-based content to more complex formats that include images, audio, and video. This transition requires more advanced detection methods capable of processing and integrating various types of data effectively. Multimodal fake news detection addresses this challenge by integrating information from different modalities, such as text and images, to improve accuracy. Since most social media content today includes both visual and textual elements, using them in combination allows models to detect inconsistencies and patterns that might not be evident when analyzing a single modality. For example, a misleading caption may be associated with a manipulated image, and understanding the relationship between the two is essential for accurate detection. Despite its potential, effective multimodal fake news detection remains in its early stages. While many studies have focused on single-modality detection, combining different data types each with unique structures and dimensions poses technical challenges. The core difficulty involves developing robust fusion strategies that can meaningfully combine information from different modalities. This survey paper focuses specifically on deep learning (DL) methods for multimodal fake news detection on social media. It reviews key works in the field, highlighting the deep learning techniques used, the data types analyzed (with a focus on text and images), and the fusion mechanisms employed. The paper also discusses major limitations in current state-of-the-art approaches.

Keywords: Fake News; Machine Learning; Deep Learning; Natural Language Processing; Computer Vision.

1 | Introduction

In today's highly connected world, ideas spread rapidly, particularly through widespread and easily accessible social media platforms. These platforms allow users to produce and share content whether in the form of text, video, audio, or images freely and without supervision. While this freedom brings many positive aspects, it has also contributed to the widespread dissemination of low-quality and misleading information, commonly referred to as fake news. Fake news is often deliberately spread by malicious actors aiming to manipulate public emotions, influence opinions, sow confusion, damage reputations, or profit from misinformation [1, 2]. It can appear in different forms: misinformation, defined as inaccurate information shared without

Corresponding Author: Mahinda.zidan@fue.edu.eg

malicious intent; disinformation, which is intentionally false information intended to mislead; and malinformation, which involves real information used maliciously to cause harm [2]. Fake news has significant impact on the social, political, economic, and personal domains [3]. On an individual level, false rumors can target innocent people, subjecting them to harassment, threats, and defamation that can lead to serious consequences in real life. In the healthcare industry, the rising dependence on the internet for medical information makes the spread of fake health news particularly dangerous, since it may cause public harm and mistrust. Economically, fake news and rumors can damage corporate reputations, manipulate markets, and influence consumer behavior, enabling unethical commercial practices. Politically, Fake news has been demonstrated to have a major impact on democratic processes and public opinion, which raises serious concerns about its effects on political integrity and stability.

The research community has focused on creating efficient detection and prevention techniques in response to increasing concerns about the spread of false information and misinformation online. One of the most promising approaches is deep learning, a subfield of machine learning that can handle large, complex datasets [4]. Deep learning has transformed fields such as natural language processing (NLP) [5] and computer vision [6], enabling the creation of powerful models that can analyze and detect fake news with increased accuracy.

Initially, studies on fake news detection focused primarily on textual content [7]. While textual analysis provides valuable insights and is an important component of misinformation detection, it has become more evident that it alone is insufficient. Online posts and articles often contain multiple modalities such as images, audio, and video through which misinformation can also be transmitted. Consequently, effective detection must go beyond text to incorporate multimodal analysis, which allows a more comprehensive understanding of misleading content.

Multimodal fake news detection, which combine data from various sources (e.g., text, images, and audio), has emerged as a critical area of research [8]. Deep learning models that can extract and fuse patterns from different data types have demonstrated significant potential in identifying both fake and real content [9]. This paper reviews recent research on multimodal fake news detection, with a specific focus on the fusion of text and image data. It also explores the fusion strategies proposed in existing studies for combining these modalities to improve detection accuracy and effectiveness.

The rest of this paper is structured in the following order: Section 2 briefly discusses A survey of deep learning techniques and general architecture of a multimodal fake news. In section 3, reviews recent representative models in the field, highlighting their approach and fusion strategies. Section 4 describes the experimental datasets commonly used for benchmarking multimodal fake news detection. Section 5 discusses the evaluation metrics used in fake news detection. Finally, section 6 concludes the paper and Future direction.

2 |Deep Learning Techniques For Multimodal Fake News Detection

Multimodal approaches which combine text and image data, improve fake news detection by utilizing the information each modality provides, especially on social media where both are present. Techniques such as feature fusion are commonly used to improve model accuracy. The general architecture of multimodal fake news detection consists of four stage (1) Multimodal input (text, image), (2) Feature extraction into separate vectors, (3) Fusion of those vectors into a unified representation, and (4) classification into “Fake” or “Real,” as illustrated in Figure 1. This section discusses the main algorithms used in multimodal fake-news detection.

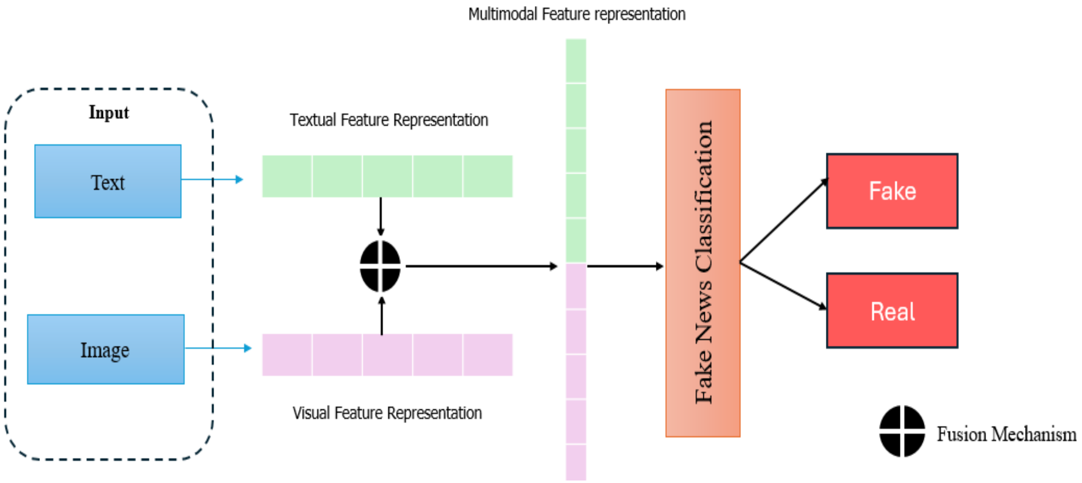
Multimodal Feature representation
Input
Text
Image
Textual Feature Representation
Visual Feature Representation
Fake News Classification
Fake
Real
Fusion Mechanism

Figure 1. The General architecture of Multimodal Fake news detection.

2.1 | Feature Extraction

Feature extraction requires generating combinations of variables to address these issues while maintaining the precision of data representation. This section is organized into two subsections: textual features and visual features.

2.1.1 | Textual-based features

Text Features convert raw text into numerical vectors that capture semantic and syntactic meaning, enabling machines to understand text. In recent years, different approaches have been developed, which include basic representations such as one-hot encoding to powerful context-aware embeddings based on deep learning. These approaches play an important role in tasks such as fake news detection. In this section highlights key algorithms used for extracting features from textual data.

One-hot vector representation [48] is one of the earliest and simplest ways for representing words .In this method, each word is encoded as an |V| x 1 vector (where |V| is the size of the vocabulary), with all values set to zero except for a single one at the index corresponding to that word in a sorted vocabulary list. While this method is simple to implement, it is unsuitable for representing large data, as it fails to capture similarities or semantic relationships between words.

Word2Vec [47] is one of the most extensively used pre-trained word embedding models. It was developed by Google and trained using the Google News dataset. It uses a shallow neural network with a single hidden layer to generate word vector representations from a text corpus. After building a vocabulary from the input text, the model learns word embeddings that preserve semantic relationships. Similarity between words can be measured using tools such as cosine similarity. Word2Vec includes two training architectures: the Continuous Bag-of-Words (CBOW) model, which predicts a target word based on its surrounding context, and the Skip-gram model, which predicts context words given a target word.

Global Vectors for Word Representation (Glove) [10] An unsupervised learning approach that creates word vector representations by utilizing the relationships between words based on global statistical information. The training process results in linear substructures within the word vector space, obtained from combined global word-word co-occurrence statistics in a corpus. The main idea of approach is that the ratios of word-word co-occurrence probabilities can capture meaningful relationships. A co-occurrence matrix records how frequently pairs of words appear together.

BERT (Bidirectional Encoder Representation from Transformers) [11] BERT represents a significant advancement over static embedding models like Word2Vec. It provides dynamic, context-sensitive embeddings based on the surrounding words in a sentence. For instance, a word with multiple meaning such as “bank” will have different embeddings depending on the context. BERT is a pre-trained transformer-based model that captures bidirectional context and can be fine-tuned for different NLP tasks. Several improved variants have since been developed, including RoBERTa (Robustly Optimized BERT) [12] and ALBERT (A Lite BERT) [13], further pushing the boundaries of contextual language modeling.

Embeddings from Language Models (ELMo) [14] also generate context-aware word embeddings, in which the vector representation of a word changes based on its usage in a sentence. Unlike Word2Vec and GloVe, which assign a fixed vector to each word, ELMo embeddings are created using deep, bi-directional LSTM networks [15] and depend on the entire input sentence. ELMo, similar to BERT, has the ability to model complex word usage patterns and capture context more effectively.

XLNet [16] presents a new technique to unsupervised language modeling known as permutation-based language modeling. It combines the advantage of BERT's bidirectional encoding with the autoregressive capabilities of Transformer-XL. This combination enables XLNet to model dependencies across different positions more efficiently and achieve a higher performance on several NLP tasks.

2.1.2 | Visual-based features

Early computer vision tasks relied on traditional machine learning (ML) models for small datasets. As data amounts increase, the deep learning (DL) models perform well at complex visual recognition tasks on large datasets. Pre-trained models, especially models trained on ImageNet, have significant impact. This section highlights the primary deep learning algorithms used to extract visual features from image data. [17]. In this section, we focus on the most popular used deep learning algorithms for visual feature extraction to detect and learn rich representations from image data.

AlexNet represented a significant breakthrough in deep learning for image classification [17]. It is a deep convolutional neural network (CNN) of eight layers, five convolutional layers followed by three fully connected layers and uses (ReLU) activation function to addresses the vanishing gradient problem. To prevent overfitting, dropout layers are employed during training, randomly disabling connections with a frequency of 0.5. AlexNet also supports multi-GPU training, which allows for faster training and the ability to handle larger models. Despite its impressive performance, the model's accuracy significantly drops if any convolutional layers are removed, indicating the importance of its full architecture.

The Visual Geometry Group Network (VGGNet)[18] is a deeper CNN architecture that builds upon AlexNet by using smaller 3×3 convolutional filters throughout the network. VGGNet comes in two popular variants: VGG-16 and VGG-19, with 16 and 19 weight layers respectively. These models are deeper than AlexNet and are also pre-trained on the ImageNet dataset. However, training such deep models is challenging due to the vanishing gradient problem. To control model complexity and reduce parameters, VGGNet employs consistent 3×3 filters with a stride of 1. These small kernels can replicate the functionality of larger filters used in AlexNet (e.g., 11×11 and 5×5) while keeping the architecture simpler and more uniform.

GoogLeNet, also known as Inception V1 [19], introduced a novel architecture that combines multiple kernel sizes within the same layer to extract features at various spatial resolutions. This innovation addresses the challenge of selecting an appropriate kernel size, as different kernels are better suited for detecting global versus local features. Rather than deepening the network, the Inception architecture widens it by integrating parallel convolutional layers with different kernel sizes. Subsequent versions, including Inception V2 and V3, refined this approach. Xception, a later development, improves upon the Inception modules by employing depthwise separable convolutions, which enhance computational efficiency and performance.

Residual Networks (ResNet) [20] were introduced to tackle the vanishing gradient problem that limits the training of very deep networks, such as VGG. ResNet uses shortcut (or skip) connections that bypass one or more layers, forming residual blocks. This enables gradients to pass more effectively through the network during backpropagation, enabling the construction of much deeper models without degradation in performance. Among the various versions of ResNet, ResNet-50 and ResNet-101 are widely used in practice for tasks such as segmentation, classification and object detection.

2.2 | Multimodal Feature Fusion Techniques

Managing multimedia data presents an inherent challenge of working with data from different modalities while preserving the correlations between them. In the context of multimodal social media posts, understanding the relationship between text and image is essential for identifying fake posts. Four techniques are commonly employed for multimodal data fusion [21].

2.2.1 | Early fusion (Data-level Fusion)

In Early Fusion [22], features from various modalities are first extracted independently and then combined into a unified representation before classification. This method focuses on integrating features early in the process to create a true joint representation of the multimedia data. One of the main advantages of early fusion is its ability to capture strong inter-modal correlations from the start, which can lead to a richer and more informative feature space for the classifier. However, one of the main disadvantages is the complexity of aligning features from different modalities and the need for extensive preprocessing, which can be especially difficult when working with limited or imbalanced datasets.

2.2.2 | Late fusion (Decision-level Fusion)

Combining the outputs from separate classifiers, each trained on a different modality. Like early fusion, it begins with extracting unimodal features; however, the key difference lies in how it processes them. Instead of merging the features early, late fusion allows each modality to learn semantic concepts independently through dedicated models. Various ensemble techniques are then used to integrate the results from these models. One of the key advantages of this method is its flexibility can handle input data streams that differ significantly in dimensionality and sampling rate. However, this approach comes with higher computational costs, as it requires training multiple separate models. Additionally, the final integration stage introduces an extra layer of learning, making the overall process more resource intensive. A further drawback is the risk of losing meaningful correlations between modalities during the final decision-level fusion.

2.2.3 | Joint Fusion or Intermediate Fusion

Enables the model to learn a joint representation of different modalities by merging their representations within a shared hidden layer. It transforms the input data into higher-level feature representations through multiple layers, allowing fusion to occur at various stages during model training. Each layer applies a combination of linear and non-linear functions to extract specific features, progressively generating new, more abstract representations of the original multimodal input.

2.2.4 | Hybrid Fusion

Integrates the characteristics of two or more fusion strategies: early, intermediate, and late fusion.

2.3 | Deep Learning Architectures for Classification

This section highlights an overview of state-of-the-art deep learning approaches used for multimodal fake news detection. Rather than analyzing into the inner working of each algorithm, the emphasis is on how these models use multiple data modalities such as text and images for improved detection accuracy. Deep learning has become a dominant approach in machine learning due to its success in fields like NLP and text mining. These models show strong capabilities to learn abstract and contextual representations from raw data, making them ideal for understanding complex patterns in fake news detection task. Neural networks (NNs) have become widely used due to their flexibility and strong performance across tasks [23]. Several studies have applied deep learning models to detect fake news in as detailed below.

Multilayer Perceptron (MLP) [24] is a basic feedforward neural network with multiple layers of nodes, which includes an input layer, hidden layers (one or more), and an output layer. Although relatively simple in comparison to other architectures, MLPs can be effective in classification tasks when combined with other feature extraction techniques.

Convolutional Neural Networks (CNN) [25] have achieved significant attention in recent years for fake news detection. CNNs, which have traditionally been employed in computer vision. CNN can extract local patterns and features, which can be useful in determining the structures and semantics of text or images data. It also can analyze large datasets quickly and efficiently, detecting complex patterns that may indicate misinformation. Their capacity to capture local relationships and scale with data size makes them ideal for large scale fake news detection

Recurrent Neural Networks RNNs [26] perform well at handling sequential data such as text, speech, and video. RNNs can identify temporal patterns, track repetitions of misleading language, and detect stylistic inconsistencies by maintaining memory of previous inputs. Their sequential processing structure makes them effective tools for identifying linguistic trends that may indicate misinformation content.

Long short-term memory. LSTM networks [27] are a type of RNN that can learn long-term dependencies. They address the vanishing gradient problem in traditional RNNs, allowing them to store information across long sequences. In the detection of fake news, LSTMs can identify repeated linguistic patterns, inconsistencies, and semantic shifts that demonstrate misleading content. In addition, LSTMs can assist in spotting misinformation sources by examining the writing styles they tend to share.

Bidirectional LSTM (Bi-LSTM) [15] extends the capabilities of common LSTMs by processing the input sequence in both forward and backward directions. This dual-context approach allows model to obtain a deeper understanding the full scope of articles , including more complex relationships. Bi-LSTMs are particularly effective in detecting subtle linguistic features and contextual variations that unidirectional models might ignore. Their ability to predict complex patterns across both past and future contexts improves accuracy in fake news classification.

Hybrid Architectures to improve performance, researchers have developed hybrid models that combined the capabilities of multiple deep learning architectures. For example, kumari et al. [26] used a model that combines BERT for text, ResNet for images, an attention-based bilinear fusion mechanism, and a Bi-LSTM followed by an MLP classifier. This multi-layer design reflects a hybrid architecture that leverages sequence modeling and deep fusion for improved fake news detection. Xue et al. [35] used a model that integrates BERT for textual features, ResNet-50 for visual features, and BiGRU layers to capture semantic consistency. It also includes a specialized module for detecting image tampering, followed by an MLP for final classification, demonstrating a hybrid approach that combines CNNs, RNNs, and dense layers. These hybrid approaches demonstrate that using multiple architectures together each targeting different aspects of the data can lead to better performance in detecting fake news.

3 | Recent Advances in Multimodal Fake News Detection

In recent years, various deep learning models have been created to solve the issues of detecting multimodal fake news. These models combine visual and textual data using a variety of fusion and classification techniques. Below, we highlight representative approaches and their key architectural components. Table 1 summarizes the main aspects of each model, including the fusion strategy used, datasets employed, performance metrics reported, and the primary limitations identified by the authors.

Qu et al. [28] proposed a QMFND model that introduces a quantum-based approach to multimodal fake news detection, using amplitude encoding and quantum convolutional neural networks (QCNNs) within a variational quantum circuit. It processes high-dimensional text and image data with improved efficiency and robustness to noise. Tested on the Gossip and Politifact datasets, it achieved 87.9% and 84.6% accuracy, respectively. Despite promising results, QMFND faces challenges such as reliance on classical NLP, training instability due to barren plateaus, data sensitivity, and current hardware scalability limitations.

Luvembe et al. [29] Presented a Complementary Attention Fusion framework (CAF-ODNN) that enhances multimodal fake news detection by combining image captioning and a bidirectional attention mechanism to align text and image features. It introduces an optimized deep neural network for high-level representation learning and achieves strong results on four datasets (CossipCo, Politifact, Fakeddit, and Pheme), with accuracy scores up to 90%. However, its major limitation lies in high computational complexity due to deep architecture and extensive hyperparameter tuning

Jing et al. [30] Designed a MPFN (Multimodal Progressive Fusion Network) model integrates BERT for text and a combination of Swin Transformer and VGG19 for image analysis. It employs a multi-level progressive fusion strategy to capture fine-grained cross-modal relationships. Evaluated on Weibo and Twitter, it achieved 83.3% accuracy on Twitter, outperforming several prior methods. However, the model's complexity, reliance on pretrained models, and potential to miss subtle text-image connections are noted as limitations.

Hua et al. [20] The TTEC model combines BERT and ResNet with back-translation for data augmentation and contrastive learning for improved feature separation. It fuses text and image features using feature-ordered concatenation and achieved a 3.1% improvement in macro F1 score on the ReCOVery dataset. While effective, its reliance on COVID-19 data limits generalizability, and its parameter tuning could benefit from more advanced optimization techniques.

Peng et al. [31] presented a CSFND model uses a two-stage approach: an unsupervised stage for context clustering with BERT/XLNet and VGG-19/ResNet, followed by a supervised detection stage using gated fusion and local classifiers. It applies late fusion and achieves 89.5% accuracy on Weibo and 83.3% on Twitter. Limitations include reliance on K-Means clustering, potential semantic-decision space misalignment, and overall model complexity.

Chen et al. [32] introduced a CAFÉ model for detecting Fake news. Model introduces a five-stage architecture combining BERT and ResNet-34 with cross-modal alignment, ambiguity estimation via KL divergence, and adaptive fusion. It adjusts feature reliance based on ambiguity levels between modalities. Evaluated on Twitter and Weibo, it showed significant accuracy gains (up to 18.9%), but faces limitations in ambiguity estimation, modality conflicts, and generalizability across diverse datasets.

Yang et al. [33] proposed a MRAN model that uses BERT and Text-CNN for text and VGG19 for image features, followed by a relationship-aware attention network with intra- and inter-modal attention blocks to model dependencies. A fully connected classifier predicts fake news based on fused features. It achieved 85.5% on Twitter, 90.3% on Weibo, and 78.0% on Pheme. Limitations include potential imbalance in modality focus and loss of hierarchical semantics due to reliance on the final output layers of models like BERT.

Yang et al. [34] presented a MCAN model that extracts spatial and frequency-domain image features using VGG19 and CNN, along with BERT-based textual features. It employs a novel multi-layer co-attention fusion mechanism to model inter-modal relationships. Evaluated on Twitter and Weibo, it achieved 80.9% and 89.9% accuracy, respectively. However, its high computational cost, sensitivity to noise, and limited generalizability present challenges for real-world deployment.

Xue et al. [35] introduced a MCNN model that combines BERT and BiGRU for text, ResNet-50 and BiGRU for image features, and includes a visual tampering module using ELA and ResNet-50 to detect image manipulation. A similarity measurement module aligns modalities before final fusion. It achieved high accuracy across multiple datasets 96.3% (Yangdataset), 94.7% (Weibo), 78.4% (Twitter), and 88.4% (Politifact). Its main drawback is the model's high computational cost, which may lead to overfitting in some scenarios.

Kumari et al. [26] presented a framework that uses an attention-based stacked BiLSTM for textual features and an attention-based CNN-RNN architecture for visual features, with Multimodal Factorized Bilinear (MFB) Pooling for feature fusion. An MLP classifier is used for final prediction. It achieved 88.3% accuracy on Twitter and 89.23% on Weibo. Despite strong results, it struggles with longer text inputs and capturing fine-grained semantic correlations between modalities, leading to occasional misclassifications.

Wang et al. [36] presented FMFN (Fine-Grained Multimodal Fusion Network) model enhances feature extraction by combining RoBERTa for text and CNN for image processing, using a scaled dot-product attention mechanism to fuse word embeddings with diverse visual feature vectors. This approach captures fine-grained interdependencies between modalities and achieved an accuracy of 88.5%. However, its evaluation on a single dataset limits conclusions about its generalizability to other domains or real-world scenarios.

Table 1. Summary of Recent Approaches in Multimodal Fake News Detection.

Ref.	Method	Data Types	Datasets	Fusion	Results	Limitation
[28]	QMFND (Quantum Encoding, VQC, QCNN)	Text + Images	Gossip and Politifact	Early Fusion (Concatenation via Amplitude Encoding)	87.9% and 84.6 %	Complexity of quantum circuits, dependence on classical processing, Barren plateau issue, reliance on classical NLP, limited training data, quantum noise, scalability.
[29]	CAF-ODNN (Complementary Attention Fusion with Optimized Deep Neural Network)	Text + Images	GossipCo, Politifact, Fakeddit and Pheme	Early Fusion (Attention Mechanism)	86.3 %, 88.9 %, 90% and 87.9 %	Computational complexity
[30]	MPFN (Multimodal Progressive Fusion Network)	Text + image	Weibo, Twitter	Hybrid Fusion (progressive fusion)	83.8% and 83.3%	May miss cross-modal connections; bias from pretrained models; high complexity
[20]	BERT + ResNet with Contrastive Learning	Text + image	ReCOVERY	Early Fusion (Feature-ordered Concatenation)	80.5% macroF1	Limited topic diversity (COVID-19 only), and basic parameter optimization method
[31]	Two-stage architecture: Unsupervised Context Learning + Supervised Detection BERT/XLNet for text, VGG-19/ResNet for images Gated fusion module Multiple local classifiers based on contextual similarity	Text + image	Weibo and Twitter	(Late Fusion) gated fusion mechanism	89.5% (Weibo), 83.3% (Twitter)	Inconsistencies between semantic and decision spaces, basic K-Means clustering, complex processing
[32]	BERT + Resnet	Text + image	Twitter and Weibo	Hybrid Fusion (Cross attention mechanism)	80.6% (Twitter), 84.0% (Weibo)	- Difficulty quantifying cross-modal ambiguity - Complexity may cause overfitting - Limited testing on diverse datasets
[33]	BERT, TEXT CNN, VGG16	Text + image	Weibo, Twitter and Pheme	Hybrid Fusion (Cross attention mechanism)	90.3% (Weibo), 85.5 % (Twitter) , 78% (Pheme)	- Difficulty capturing full intra/inter-modal relationships - Loss of hierarchical semantics - Imbalanced attention to modalities
[34]	BERT, VGG19, Co-Attention Mechanism	Text + Image	Twitter, Weibo	Early Fusion (Co-Attention Mechanism)	Twitter 80.9%, Weibo 89.9 %	Complexity of Co-Attention Mechanism High Computational Cost Sensitivity to Noise Limited Generalization.

[35]	BERT, Resnet50, BiGRU , Attention	Text + image	Yang dataset, Weibo, Twitter, Politifact	Early Fusion (Attention Mechanism)	96.3%, 94.7%, 78.4%, 88.4% respectively	Complexity may cause overfitting.
[26]	Attention-based BiLSTM + CNN-RNN + MFB + MLP	Text + image	Twitter, Weibo	joint/intermediate MFB Pooling	88.3% (Twitter), 89.23% (Weibo)	Struggles with long texts and weak inter-modal correlation handling
[36]	FMFN (Fine-Grained Multimodal Fusion Network)	Text+image	Weibo	Early Fusion Scaled Dot-Product Attention	88.5%	Limited generalizability due to evaluation on only one dataset

4 |Multimodal Datasets For Fake News Detection

Weibo dataset is a collection of posts from Sina Weibo, a major Chinese microblogging platform similar to Twitter. It contains diverse content types, including text, images, videos, and links. Commonly used in research areas like sentiment analysis, it offers valuable insights into social media behavior in China. This paper references two versions of the dataset: Weibo A [37] and Weibo B [38].

Fakeddit [39] is a large-scale multimodal dataset with over 1 million samples referring to various forms of fake news. Each sample contains a submission title, an image, comments, and extra metadata like rankings and comment counts. The dataset goes through multiple review actions and is labeled using distant supervision for classification tasks in 2-way, 3-way, or 6-way categories, making it appropriate for training and testing false news detection models.

Twitter dataset [40] Corpus was used in the MediaEval 2015 and 2016 workshops for the "Verifying Multimedia Use" task. In MediaEval 2015, the dataset contained 11 events with a training set of 5,008 real and 6,840 fake tweets, and a test set of 1,217 real and 2,564 fake tweets. Some rumor tweets that were originally included were later removed, which had a little impact on the fake tweet count. In MediaEval 2016, the 2015 training and test sets were combined into a single training set, while a new test set of 1,107 real and 1,121 fake tweets was introduced. This dataset has become a standard for multimodal fake news detection tasks on social media.

FakeNewsNet [41] It is a publicly available dataset created mainly to enable research into false news identification. It contains a wide range of data, including textual content, photos, and social media engagement metrics such as retweets and likes. The dataset consists of two main sub-datasets:

- PolitiFact: Contains fact-checked news articles from PolitiFact, covering political content published between May 2002 and July 2018.
- GossipCop: Focuses on entertainment-related news, with credibility scores ranging from 0 (completely false) to 10 (completely true), sourced from the GossipCop website.

Each sub-dataset includes both real and fake news articles with appropriate labels. FakeNewsNet is a valuable resource for developing and benchmarking models in NLP, ML, and social media analysis for fake news detection.

BuzzFeed [46] News dataset covers almost 200,000 articles published between 2014 and 2018 on a wide range of topics including politics, entertainment, and technology. Each article has metadata such as title, publication date, URL, author, complete text, photos, and social media interaction metrics (for example, likes, shares, and comments). The dataset is frequently utilized in research, including topic modeling, sentiment analysis, and false news identification.

NewsBags dataset [45] consists of 215,000 news articles, with 15,000 fake news items sourced from The Onion and 200,000 real news articles from The Wall Street Journal. Due to class imbalance, a revised version called NewsBag++ was created, increasing the number of fake news items to 389,000. An additional test set includes 11,000 real and 18,000 fake articles. NewsBag is useful for binary fake news classification tasks and evaluating model performance under class imbalance conditions.

The Yang dataset [42] consists of 20,015 news articles, including 11,941 fake and 8,074 authentic news items. The fake news samples were collected from over 240 unreliable websites, while the real news came from reputable sources like The New York Times, The Washington Post, and others. Each news item includes multiple fields such as title, text, image, author, and website, making it a rich multimodal resource for fake news detection research.

ReCOVery [44] dataset includes 2,029 news articles related to the COVID-19 pandemic, published between January and May 2020. Each news item is labeled as real or fake based on source credibility. The dataset is multimodal, providing information such as text, image, source, publication time, and author details making it ideal for studying misinformation during global crises.

The PHEME dataset [43] is a public dataset designed for rumour detection and veracity classification on social media. It includes tweets from nine major events. It is divided into four sub-datasets: Rumours: Tweets labeled as true or false; Non-rumours: Tweets not linked to rumours; Thread structure: Information about tweet threads (e.g., number of replies/retweets).; Stance: Annotations of whether tweets support, deny, or remain neutral toward a rumour. Each tweet also includes metadata like timestamp, user, and location. PHEME is especially useful for research in stance detection, rumour propagation, and event-based misinformation.

5 | Evaluation Metrics

Fake news detection is primarily formulated as a binary classification task, where the goal is to determine whether news is fake or real. Consequently, standard classification metrics are widely used to evaluate model performance. This section presented the most common evaluation metrics, along with their basic definitions.

5.1 | Confusion Matrix Terms

To define the metrics, we start with four basic terms from the confusion matrix:

- True Positives (TP): Number of samples correctly identified as positive (e.g., fake news successfully classified as fake).
- True Negatives (TN): Number of samples correctly identified as negative (e.g., real news successfully classified as real).
- False Positives (FP): Number of samples incorrectly identified as positive (e.g., real news wrongly classified as fake).
- False Negatives (FN): Number of samples incorrectly predicted as negative (e.g., fake news wrongly classified as real).

5.2 | Common Evaluation Metrics

- **Accuracy** measures the overall correctness of the model by calculating the ratio of correct predictions to total predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** calculates how many of the predicted positive samples are actually positive:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall** (True Positive Rate (TPR) or Sensitivity) measures how many of the actual positive samples were correctly predicted:

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score** is the harmonic means of precision and recall:

$$F1 - Score \ = 2 * \frac{Precision * recall}{Precision + recall}$$

- **Area Under Curve (AUC)** In order to understand AUC, it is essential to first understand the following metrics:
- **False Positive Rate (FPR)** is calculated according to the following rule:

$$FPR = \frac{FP}{TN + FP}$$

- **Receiver Operating Characteristic (ROC) Curve**: The ROC curve is generated by plotting the (TPR) against (FPR). It provides a way to evaluate the performance of a classification model at a certain threshold.

Thus, AUC summarizes the ROC curve and measures how well a model can differentiate between positive and negative classes. A model with a higher AUC is considered to have better performance.

6 | Conclusion and Future Directions

The rise of fake news alongside the rapid growth of social media presents an ongoing threat to social stability. In response, the researchers are actively working to develop effective fake news detection strategies. This survey presents a detailed analysis of fake news detection, focusing on recent and advanced techniques combining natural language processing (NLP), visual feature extraction, and deep learning (DL)techniques. DL plays an important role in this field, providing powerful tools for modeling complex patterns across different sources of information. We studied several textual and visual feature extraction methods, focusing on their importance in multimodal fake news detection. Additionally, we examined fusion approaches that combine data from several modalities, highlighting their strengths and weaknesses. Also, we summarized a variety of DL architectures commonly used in fake news detection and examined their contributions and concerns. Moreover, we discussed performance metrics used to evaluate model effectiveness and highlighted the important findings from recent experimental studies.

As a fake news evolves, particularly in multimodal formats, it indicates that this will remain an important topic of research. Developing deep learning models and novel fusion methodologies are anticipated to drive future growth in the future. We hope that study is useful resource for researchers, providing a straightforward and easily understood knowledge of the present situation, existing challenges, and interesting areas for future work in multimodal false news detection.

Acknowledgments

The authors are grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Funding

This research has no funding source.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newslett., vol. 19, no. 1, pp. 22–36, 2017.

[2] K. SANTOS-D'AMORIM and M. K. F. de Oliveira MIRANDA, "Misinformation disinformation and malinformation: clarifying the definitions and examples in disinfodemic times", Encontros Bibli: revista eletronica de biblioteconomia e ciencia da informacao, vol. 26, 2021.

[3] Kalsnes, Bente. 2018. Fake News. In Oxford Research Encyclopedia of Communication. Oxford: Oxford Research Encyclopedia.

[4] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal fake news detection," Information, vol. 13, no. 6, p. 284, Jun. 2022. [Online]. Available: https://www.mdpi.com/2078-2489/13/6/284

[5] M. D. Ibrishimova and K. F. Li, "A machine learning approach to fake news detection using knowledge verification and natural language processing," in Proc. Int. Conf. Intell. Netw. Collaborative Syst. Cham, Switzerland: Springer, 2019, pp. 223–234.

[6] D. Shen, G. Wu, and H. Suk, Deep learning in medical image analysis," Annu. Rev. Biomed. Eng., vol. 19, pp. 221–248, Jun. 2017

[7] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," 2018, arXiv:1811.00770.

[8] K. Dhruv, G. J. Singh, G. Manish, and V. Vasudeva, "MVAE: Multimodal variational autoencoder for fake news detection," in ACM World Wide Web Conference, 2019.

[9] J. Gao et al., "A survey on deep learning for multimodal data fusion," Neural Computation, vol. 32, no. 5, pp. 829–864, 2020.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proc. 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532– 1543.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in ICLR, 2019.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in NAACL, 2018, pp. 2227–2237

[15] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in Proc. IEEE Workshop Autom. Speech Recognit. Understand., Olomouc, Czech Republic, Dec. 2013, pp. 273–278, doi: 10.1109/ASRU.2013.6707742.

[16] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, volume 25, pages 1097–1105, 2012.

[18] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru and S. Satoh, "SpotFake: A Multi-modal Framework for Fake News Detection", Proceedings of the IEEE 5th International Conference on Multimedia Big Data BigMM, pp. 39-47, 2019.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015. 5.

[20] J. Hua, X. Cui, X. Li, K. Tang and P. Zhu, "Multimodal fake news detection through data augmentation-based contrastive learning", Appl. Soft Comput., vol. 136, 2023.

[21] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, Multimodal deep learning for biomedical data fusion: A review,' Briefings Bioinf., vol. 23, no. 2, Mar. 2022, Art. no. bbab569.

[22] E. F. Ayetiran and Ö. Özgöbek, "An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection," Inf. Syst., vol. 123, Jul. 2024, Art. no. 102378, doi: 10.1016/j.is.2024.102378.

[23] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," Pattern Recognit., vol. 15, no. 6, pp. 455–469, Jan. 1982, doi: 10.1016/0031- 3203(82)90024-3.

[24] S. Tufchi, A. Yadav and T. Ahmed, "A comprehensive survey of multimodal fake news detection techniques: advances challenges and opportunities", International Journal of Multimedia Information Retrieval, vol. 12, no. 2, pp. 28, 2023.

[25] Q. Li, Q. Hu, Y. Lu, Y. Yang, and J. Cheng, "Multi-level word features based on CNN for fake news detection in cultural communication," Pers. Ubiquitous Comput., vol. 24, no. 2, pp. 1–14, 2019.

[26] R. Kumari and A. Ekbal, "AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection," Expert Syst. Appl., vol. 184, Dec. 2021, Art. no. 115412.

[27] S. Deepak and B. Chitturi, "Deep neural approach to Fake-News identification,' Proc. Comput. Sci., vol. 167, pp. 2236–2243, Jan. 2020. [Online]. Available: http://www.sciencedirect. com/science/article/pii/S1877050920307420

[28] Z. Qu, Y. Meng, G. Muhammad and P. Tiwari, "QMFND: A quantum multimodal fusion-based fake news detection model for social media", Inf. Fusion, vol. 104, 2024.

[29] A. M. Luvembe, W. Li, S. Li, F. Liu and X. Wu, "CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection", Inf. Process. Manage., vol. 61, no. 3, 2024.

[30] J. Jing, H. Wu, J. Sun, X. Fang and H. Zhang, "Multimodal fake news detection via progressive fusion networks", Inf. Process. Manage., vol. 60, no. 1, Jan. 2023.

[31] L. Peng, S. Jian, Z. Kan, L. Qiao, and D. Li, "Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection," Inf. Process. Manage., vol. 61, no. 1, Jan. 2024, Art. no. 103564.

[32] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In Proceedings of the ACM Web Conference 2022. 2897–2905.

[33] H. Yang, J. Zhang, L. Zhang, X. Cheng, and Z. Hu, "MRAN: Multimodal relationship-aware attention network for fake news detection," Comput. Standards Interface, vol. 89, Apr. 2024, Art. no. 103822.

[34] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2560– 2569.

[35] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management 58, 5 (2021), 102610.

[36] J. Wang, H. Mao and H. Li, "FMFN: Fine-grained multimodal fusion networks for fake news detection", Appl. Sci., vol. 12, no. 3, 2022.

[37] Jin Z, Cao J, Guo H, Zhang Y, Luo J (2017) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on mul-timedia. MM'17. Association for Computing Machinery, New York, pp 795–816.

[38] Cao J, Sheng Q, Qi P, Zhong L, Wang Y, Zhang X (2019) False news detection on social media. arXiv. https:// doi. org/ 10. 48550/ ARXIV.1908.10818.

[39] Nakamura K, Levy S, Wang WY (2019) r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. arXiv. https://doi.org/10.48550/ARXIV.1911.03854.

[40] C. Boididou et al., "Verifying multimedia use at MediaEval 2015", Proc. MediaEval, 2015.

[41] Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2019b) FakeNews Net: a data repository with news content, social context and spatial temporal information for studying fake news on social media.

[42] Yang Y, Zheng L, Zhang J, Cui Q, Li Z, Yu PS (2018) TI-CNN: convolutional neural networks for fake news detection. arXiv.

[43] Zubiaga A, Liakata M, Procter R, Hoi GWS, Tolmie P (2016) Analys-ing how people orient to and spread rumors in social media by looking at conversational threads. PLOS ONE 11(3):0150989. https://doi.org/10.1371/journal.pone.0150989.

[44] Zhou X, Mulay A, Ferrara E, Zafarani R (2020a) Recovery: a multimodal repository for COVID-19 news credibility research. In: CIKM'20. Association for Computing Machinery, New York, pp 3205–3212. https://doi.org/10.1145/3340531.3412880.

[45] S. Jindal, R. Sood, R. Singh, M. Vatsa and T. Chakraborty, "Newsbag: A multimodal benchmark dataset for fake news detection", CEUR Workshop Proc, vol. 2560, pp. 138-145, 2020, February.

[46] C. Silverman, L. Strapagiel, H. Shaban, E. Hall, and J. Singer-Vine, "Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate," oct 2016. [Online]. Available: https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis.

[47] Zhang F (2022) A hybrid structured deep neural network with Word2Vec for construction accident causes classification. Int J Constr Manag 22(6):1120–1140

[48] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang, "Cross-modal contrastive learning for multimodal fake news detection," in Proc. 31st ACM Int. Conf. Multimedia, Oct. 2023, pp. 5696–5704.